番組コンテンツに対するユーザ入力情報からの時系列キーワード抽出に関する一考察

大黒　泰平† 加藤　友規‡ 土居　清之‡ 亀山　渉†

早稲田大学大学院国際情報通信研究科† 日本テレビ放送網株式会社‡

1. はじめに

テレビ放送がディジタル化するとともに、大容量で低価格なハードディスクを利用した「蓄積型受信機」の普及が予想される。蓄積型受信機では、視聴者（ユーザ）が欲するコンテンツを時間や形態を問わずに自由に視聴できる環境が望まれている。そのためには視聴補助情報（メタデータ）の制作と提供が不可欠である。

本稿では、視聴者が番組に対して抱いた感情や感想を基に、放送局が番組に対してメタデータを付与することを最終目的とするキーワード抽出手法を検証する。具体的には、番組のチャットルームを利用し、視聴者の番組の特定部分に対する発言を時間軸に沿ってキーワードとして定義し、視聴者の番組に対する情報（感情・感想）のリアルタイム取得について検討を行ったので、その結果を報告する。

2. 実験システム

2.1 概要

ユーザの番組コンテンツに対する感想を調査するために、ユーザ間で番組コンテンツに関して PC でチャットを行える環境を構築した。この環境において、ユーザがチャットで入力するキーワードの発生傾向を調査するための被験者実験を行った。実験の環境は以下の通りである。

- ●被験者　　：10 名(男性 6 名、女性 4 名)
- ●番組視聴時間　：約 57 分
- ●番組内容　　：バラエティー番組

なお、番組コンテンツに対する発言を多く取得するため、番組コンテンツに全く関係のない発言は避けるよう教示した。被験者には、TV 画面で番組コンテンツを視聴しながら、PC 端末画面上に表示されたチャットページ上で自由に会話を行ってもらい、データを取得した。

実験の結果、57 分間で約 190 回確認の発言が確認された。それらの発言の中から、以下に定義したキーワードがどの程度含まれているのかを調査した。

2.2 キーワードの定義

ユーザが発言した全ての文字列の中から、キーワードであると考えられる語句の 3 分毎の発生頻度を調べた。

なお、定義したキーワードは、

① 番組内で発言された、または字幕として表示された、番組そのものと関係の深い語句
② 「難しい」や「?」等、ユーザの疑問を表す語句
③ 「!」の等、ユーザの強い感情を表す語句
④ 「CM」というキーワード
⑤ 「CM」と関係の深いキーワード

の 5 種類である。

これらのキーワードを定義した理由について、①は、ユーザが番組コンテンツのどの部分に興味を持っているのかを測定する指標になると考えたため定義した。②・③は、ユーザが番組コンテンツに対してどのような感情、感想を持っているのかを測定する指標となると考えたため定義した。④・⑤は、ユーザがどのようなジャンルの CM に対して興味を持っているのかを測定する指標になると考えたため定義した。

図 1 に 3 分毎のユーザの総発言回数を、図 2 にキーワードの発生頻度を示す。また、番組コンテンツ内において起きた主なイベントについて、時間軸に記載した。

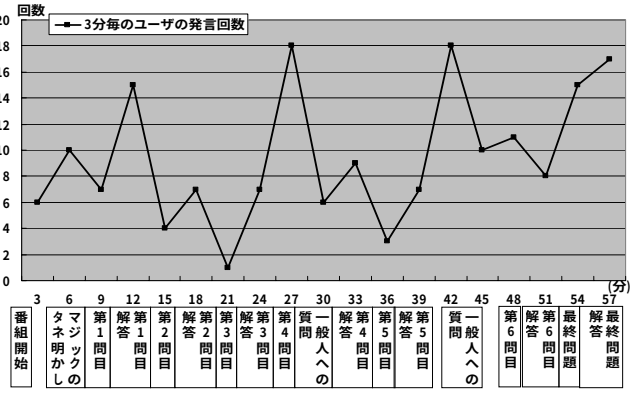

図 1.3 分毎のユーザの発言回数

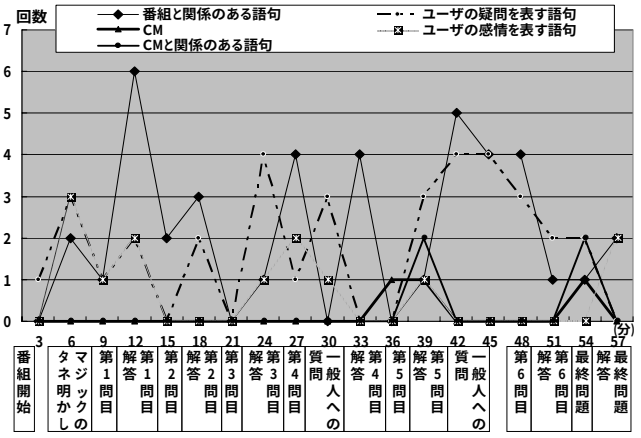

図 2.各種キーワードの発生回数

A Consideration on Extracting Time-Based Keywords from a TV Program

† Taihei DAIKOKU, E-mail:taihei_d@fuji.waseda.jp
Wataru KAMEYAMA
GITS Waseda University
‡ Tomonori KATO
Kiyoyuki DOI
Nippon Television Network Corporation

3. 実験結果

実験の結果、以下のことが明らかとなった。

① 番組が開始されてから、発言が安定して書き込まれるようになるまで多少の時間(約 6 分)がかかる。[図 1]

② ユーザの発言は常に一定の回数書き続けられているわけではない。傾向として、問題が提起されている場面では少なく、問題を解答している場面では多く発言されている。[図 1]

③ 番組と関係の深い語句が多く発言されるのは、問題が提起されている時ではなく、問題が解答されている時である。[図 2]

④ 12 分の時点では問題文に関係のある語句とユーザの感情を表す語句共に多く発言されている。しかし、42 分の時点では問題文に関係のある語句は多く発言されているが、ユーザの感情を表す語句は発言されていない。このことから、12 分の時点の問題は、ユーザに対して感情的に訴えかけるであったが、42 分の時点の問題はユーザに対して感情的に訴えかけるものではなかった、などと考えることができる。このように、1 つのキーワードだけではわからないことも複数のキーワードの発言頻度を組み合わせることにより情報が抽出できる可能性もある。[図 2]

⑤ 全体的にユーザの疑問を表すキーワードが多く発言されるのは問題が提起されている時ではなく、問題が解答される時である。このことから、ユーザが問題に対して疑問を持つのは、問題が提起された時ではなく、問題が解答された時である、などと考えることができる。[図 2]

⑥ 番組が CM に入った後は、CM というキーワードが多く発言される傾向がある。[図 2]

このような考察を繰り返すことにより、ユーザが番組のどの時点において盛り上がりを見せているのか、どのような場面に対して興味を持っているのかなど、番組コンテンツ内の特定の場面に対する盛り上がり具合を調査することも可能になるものと考える。

4. 今後の課題

今回は、ユーザからのキーワードの発言傾向を時系列に沿って見ることに主眼を置いた。しかし、番組コンテンツの内容によって発言されるキーワードの種類は変わってくるものと考えられる。異なるジャンルの番組コンテンツにおいても同様の調査を行う必要がある。

また、ユーザの特定の感想を反映しているキーワードは一様の語句だけで表現されるわけではない。それらの一連の語彙をシソーラスとして一元的に定義[2]していくべきである。そして、シソーラスを利用した抽出精度の高いキーワード抽出を行う必要があり、図 3 に示すような方法を検討する必要がある。

① ユーザが入力した全文字列に対して Cha-Sen[3]を用いて形態要素解析を行う。

② ①で解析された品詞のうち、明らかにキーワードとはなり得ないと考えられる語を削除リストとしてパターン化し、そのパターンに当てはまる語句を削除する。

③ ②において削除されなかった語を品詞毎に分類し、品詞リストとしてパターン化する。

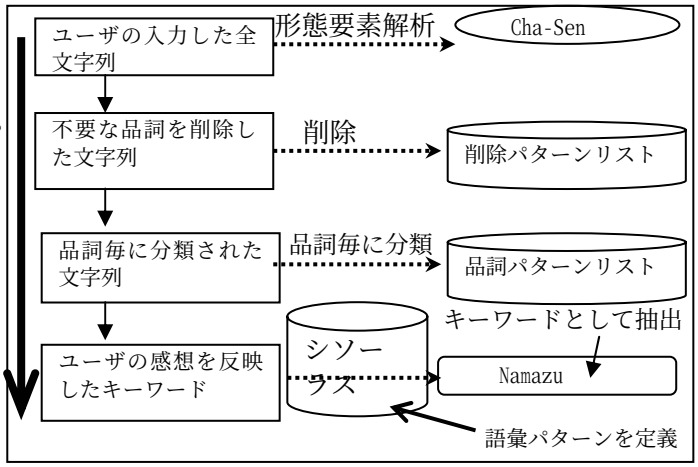

図 3.キーワードの抽出手法

④ ユーザの感情を表すと考えられる語句をシソーラスとして定義し、パターン化する。

⑤ ④で定義されたシソーラスの語彙を元に、Namazu[4]を用いて③の品詞リストに対してキーワードの抽出を行う。

⑥ ユーザの特定の感想を反映したキーワードが Namazu によって抽出される。

上記処理を行うことにより、段階的にユーザが入力した文字列がフィルタリングされ、最終的にユーザの感情や感想を表す文字列をキーワード単位で抽出することができるものと考える。

しかし、ユーザからの情報をキーワード単位で抽出するだけでは再利用可能な有益な情報であるとは言い難い。よって、抽出したキーワードを再利用可能なメタデータに再構築することも今後の課題である。

5. まとめ

本稿では、番組コンテンツに対して、ユーザ入力情報から時系列キーワードの発生傾向を調査した。その結果、ユーザが難しいと感じていると考えられる場面では疑問を表すキーワードが多く抽出されるなど、キーワードの発生傾向にある法則性が見られることが明らかになった。

この様な特性を用い、ユーザが番組コンテンツに対して抱く感情を時系列に沿って定義することができれば、ユーザの嗜好情報をリアルタイムに反映した情報を得ることが可能になると考える。

参考文献

[1]石山芳彦，細川博之，“新聞記事からの組織情報の抽出”，第 65 回情報処理学会全国大会、3Z-01，2003

[2]川島貴広，石川勉，“言葉の意味に関する類似性判別能力における概念ベースとシソーラスとの性能比較”，第 65 回情報処理学会全国大会，2M-1，2003

[3]Cha-Sen，http://chasen.aist-nara.ac.jp/index.html.ja

[4]Namazu，http://www.namazu.org/